



TopicView: Visually Comparing Topic Models of Text Collections

November 7, 2011

Patricia Crossno, Andrew Wilson, Timothy Shead,
Daniel Dunlavy

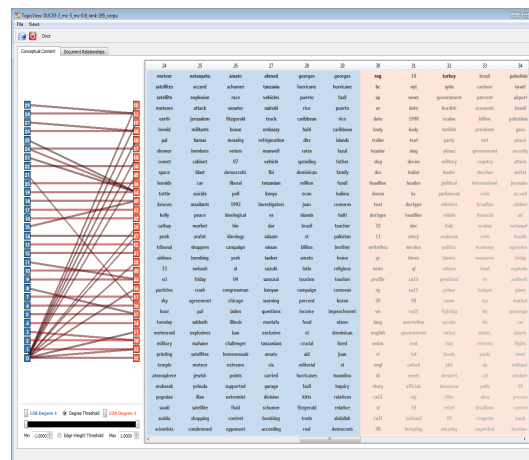
Sandia National Laboratories

Modeling Text Data

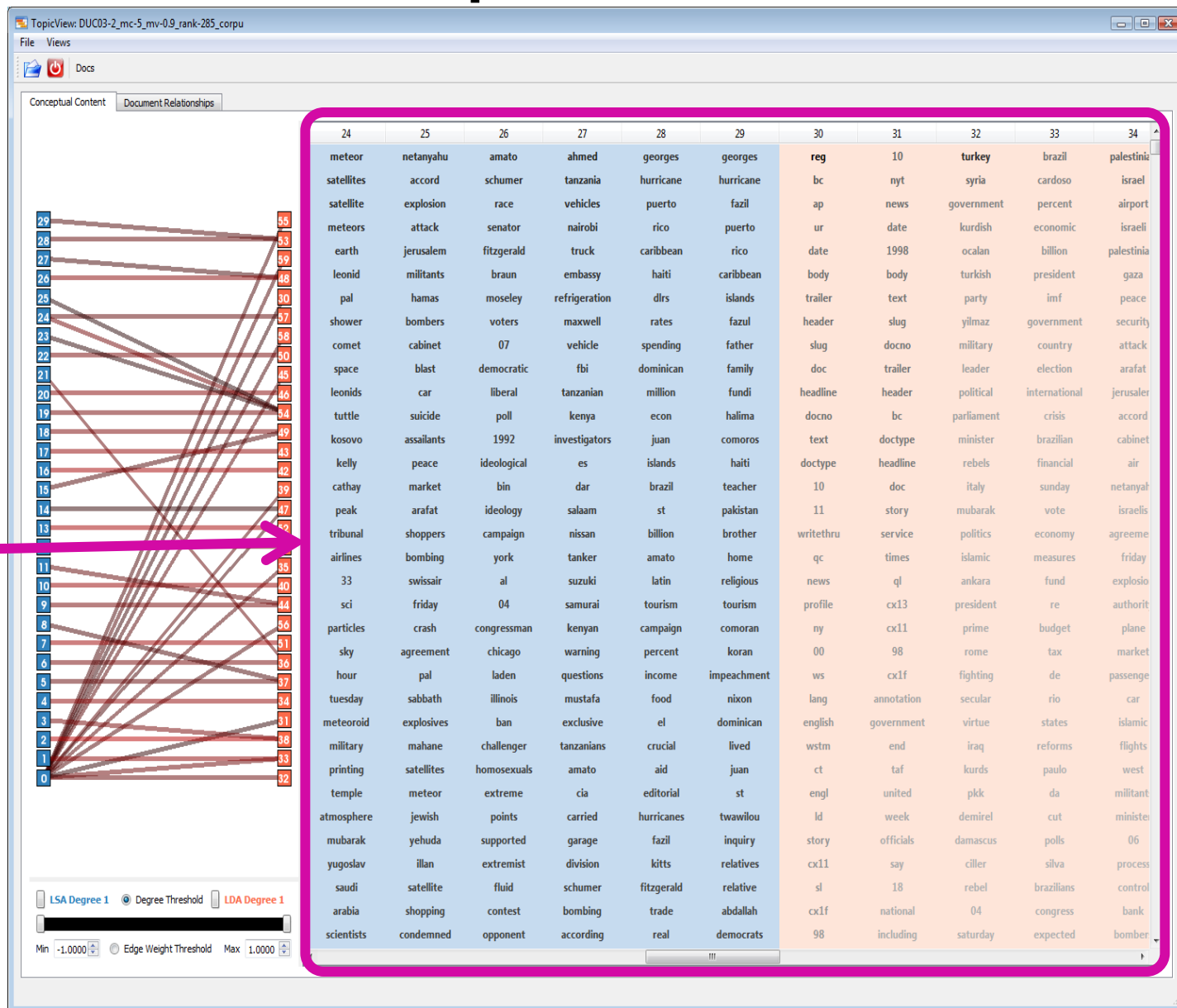
- Latent Semantic Analysis (LSA) vs Latent Dirichlet Allocation (LDA)
- Similarities
 - Bag-of-words modeling
 - Transform text to term-document frequency matrices
 - User-defined # of dimensions
 - Produce weighted term lists for each concept/topic
 - Produce topic weights for each documents
 - Results used to compute document relationship measures
- Differences
 - LSA: truncated singular value decomposition (SVD) -> correlations (-1 to 1)
 - LDA: Bayesian model -> probabilities (0 to 1)
 - Output quantities have different ranges and meanings
- Direct numeric comparison not meaningful

Comparing LSA and LDA

- Focus on how models used in applications
- Conceptual content
 - Topic models
 - Labels
- Document relationships
 - Scatter plots
 - Graphs
 - Landscapes
- TopicView application
 - Visually compare and interactively explore models
 - Tabbed panels (Conceptual Content & Document Relationships)
 - Linked views
 - Built using Titan Informatics Toolkit



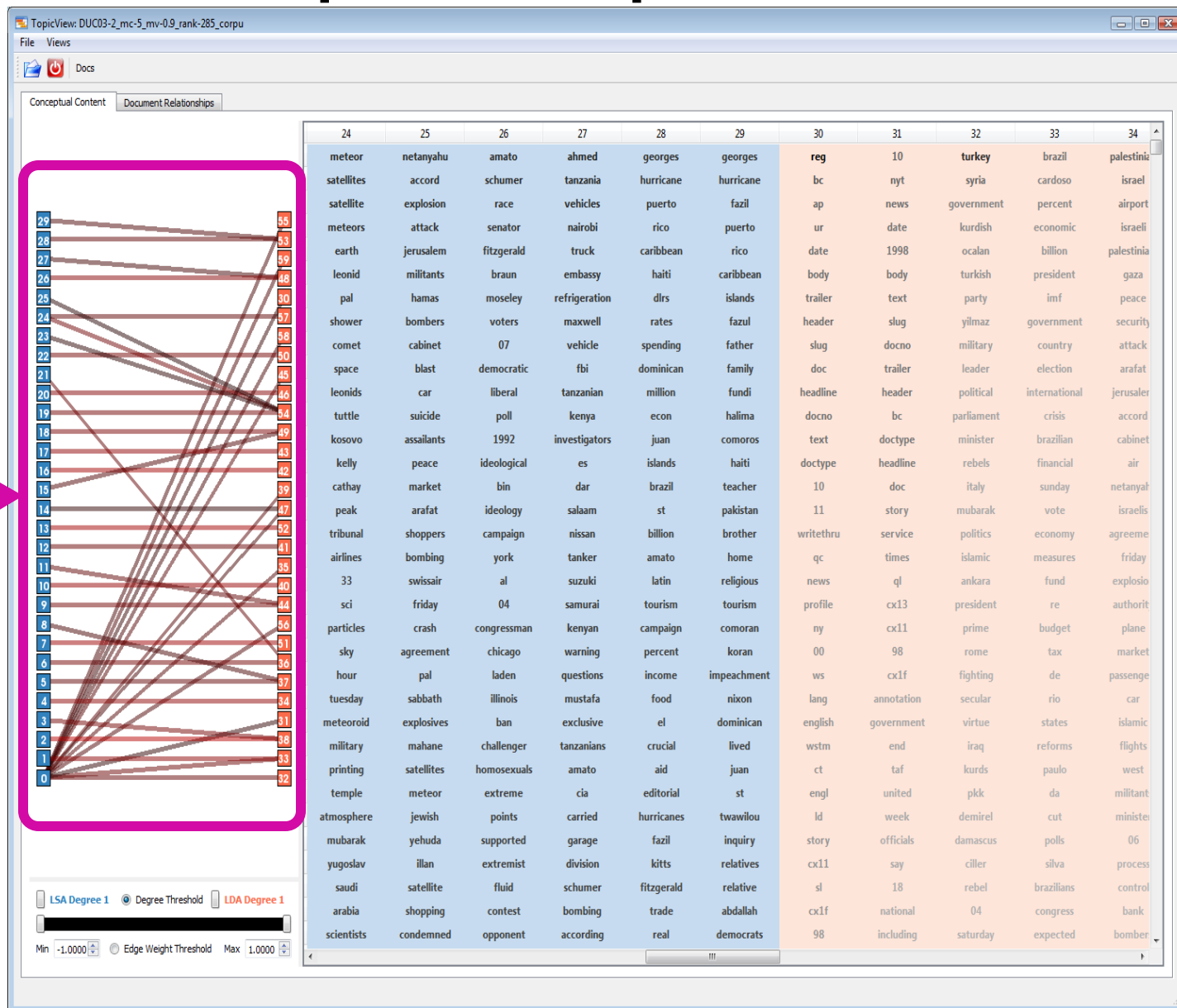
Term Topic Table

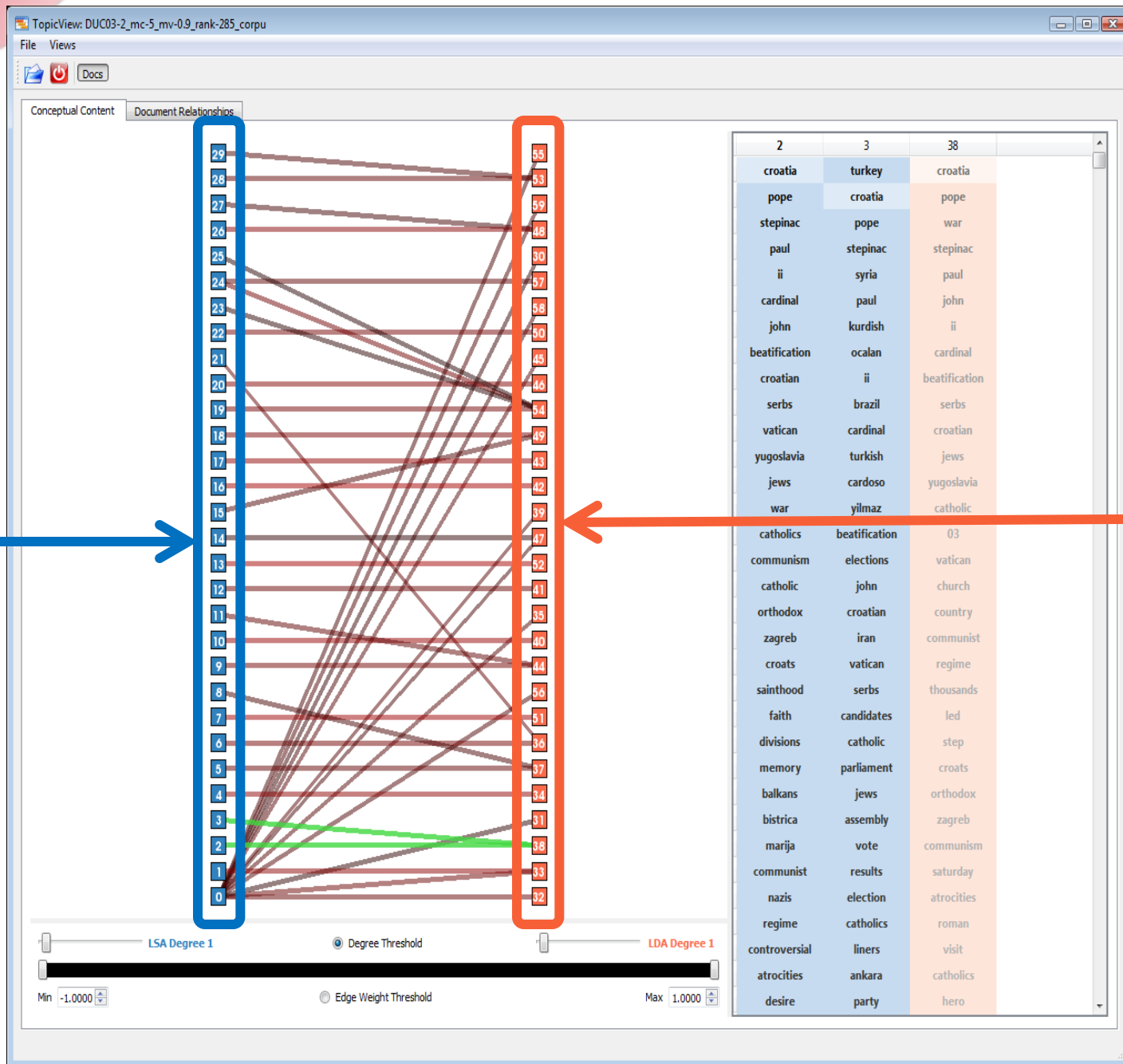


Detailed
Conceptual
Similarity

Bipartite Graph

High-level
Conceptual
Similarity

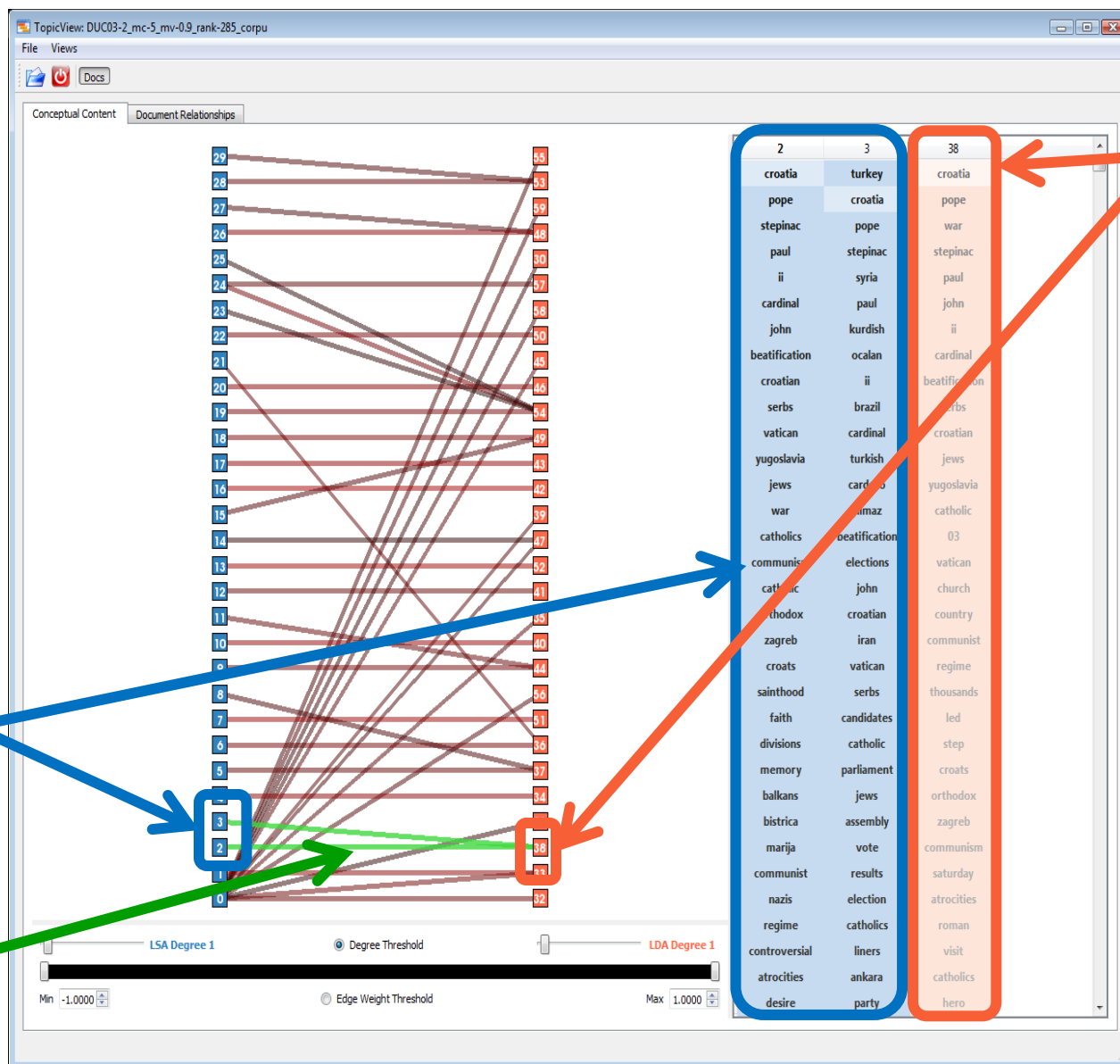




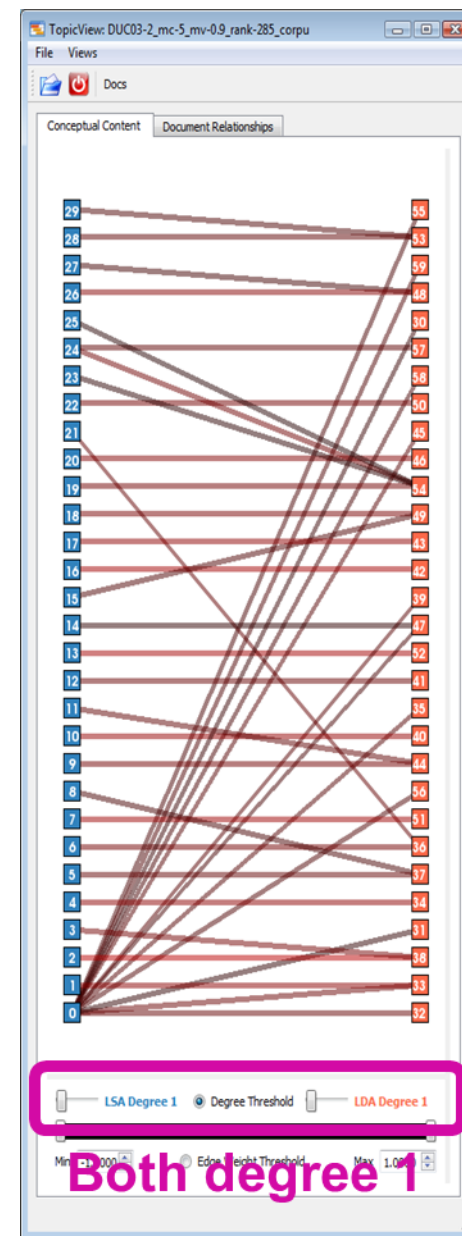
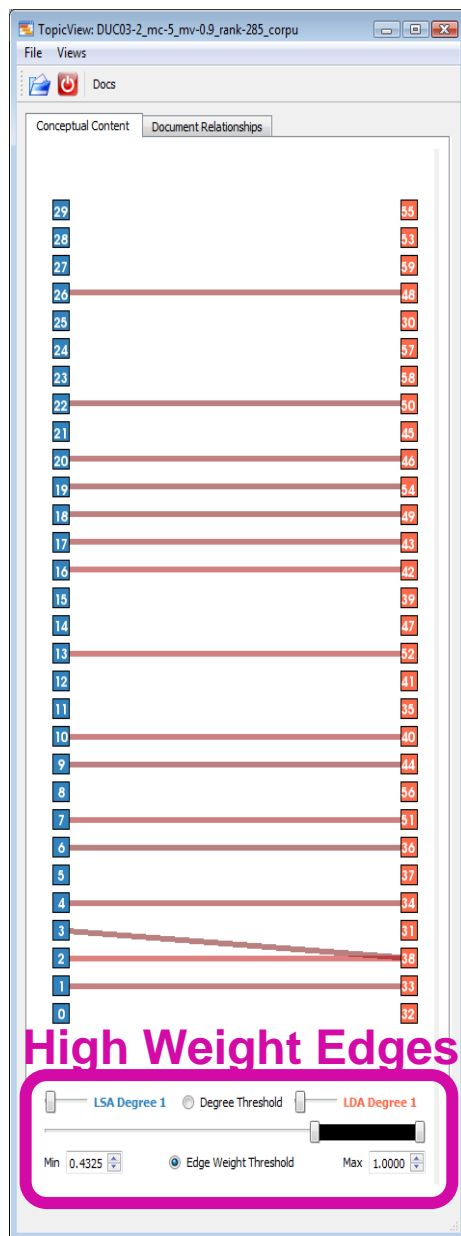
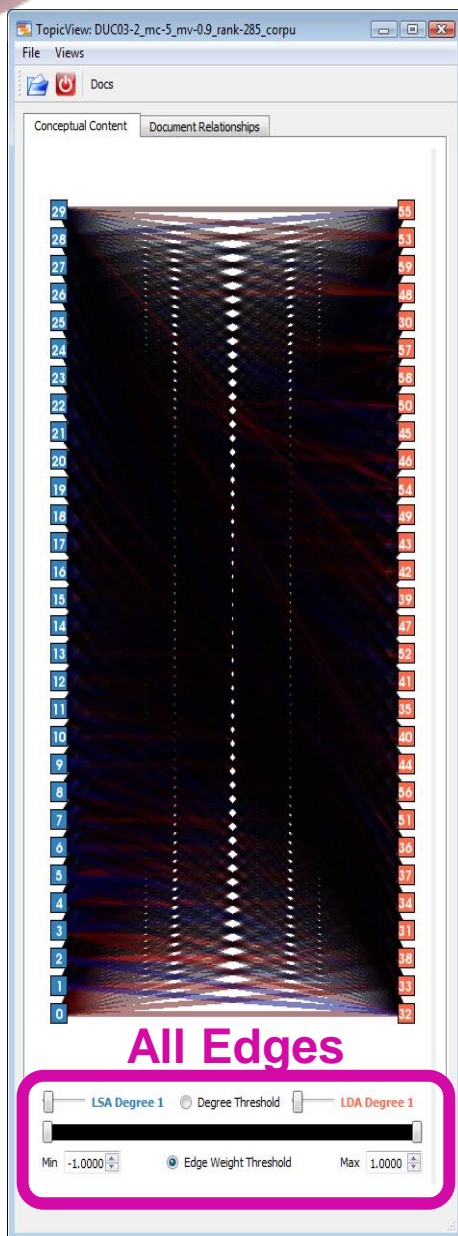
LSA
Concepts

LDA
Topics

Linked Selection

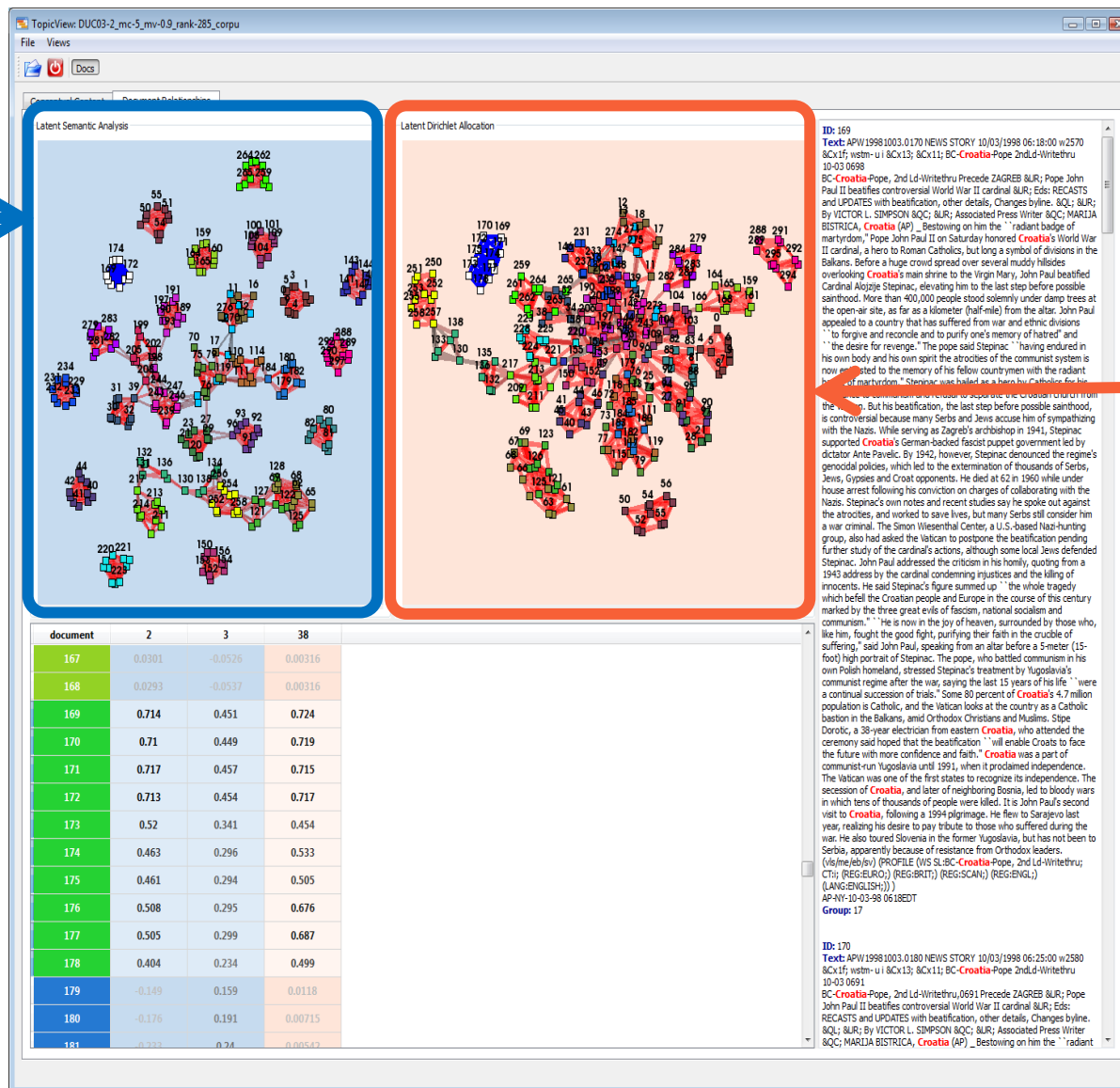


Edge Display Controls



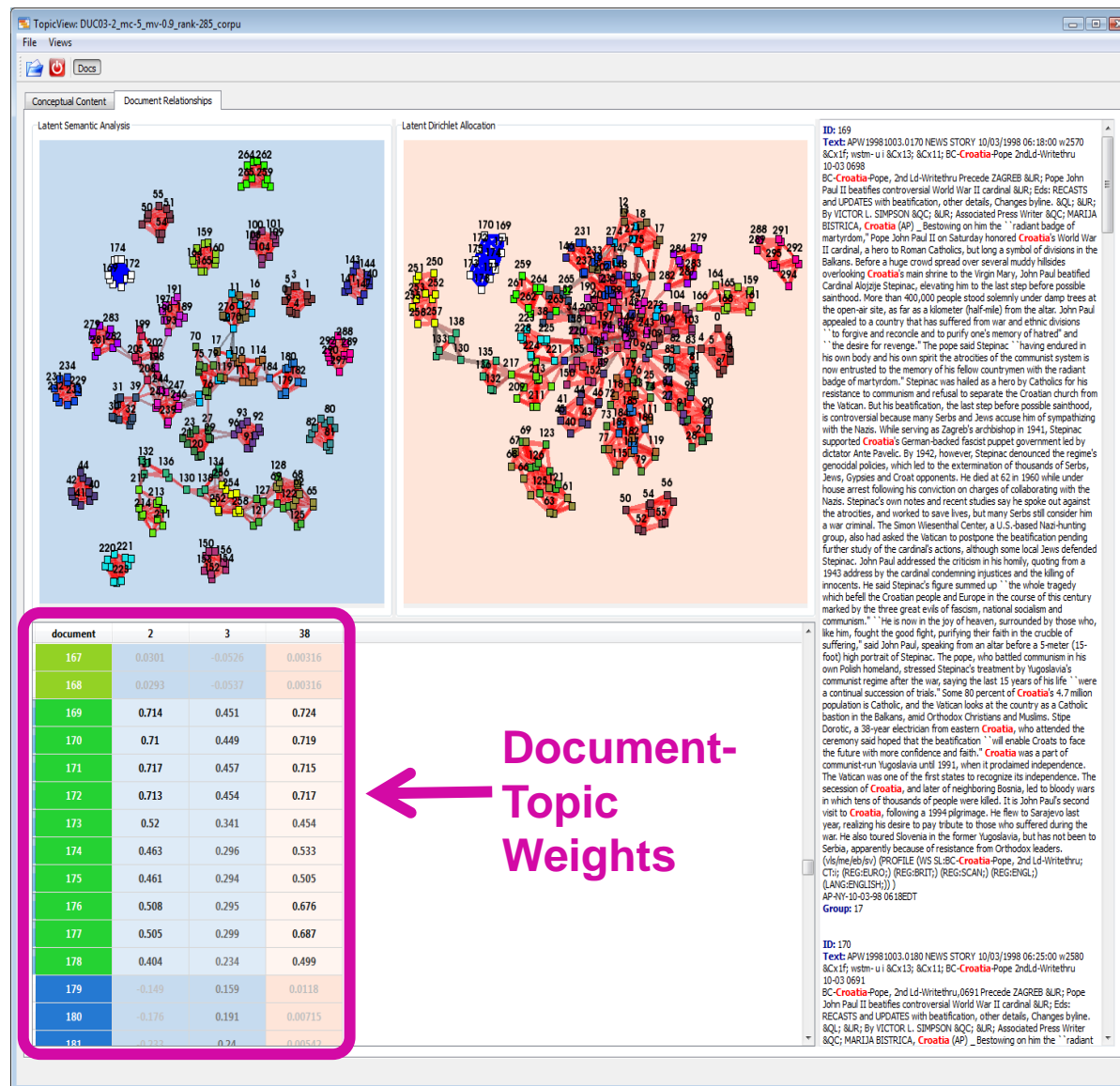
Document Relationship Graphs

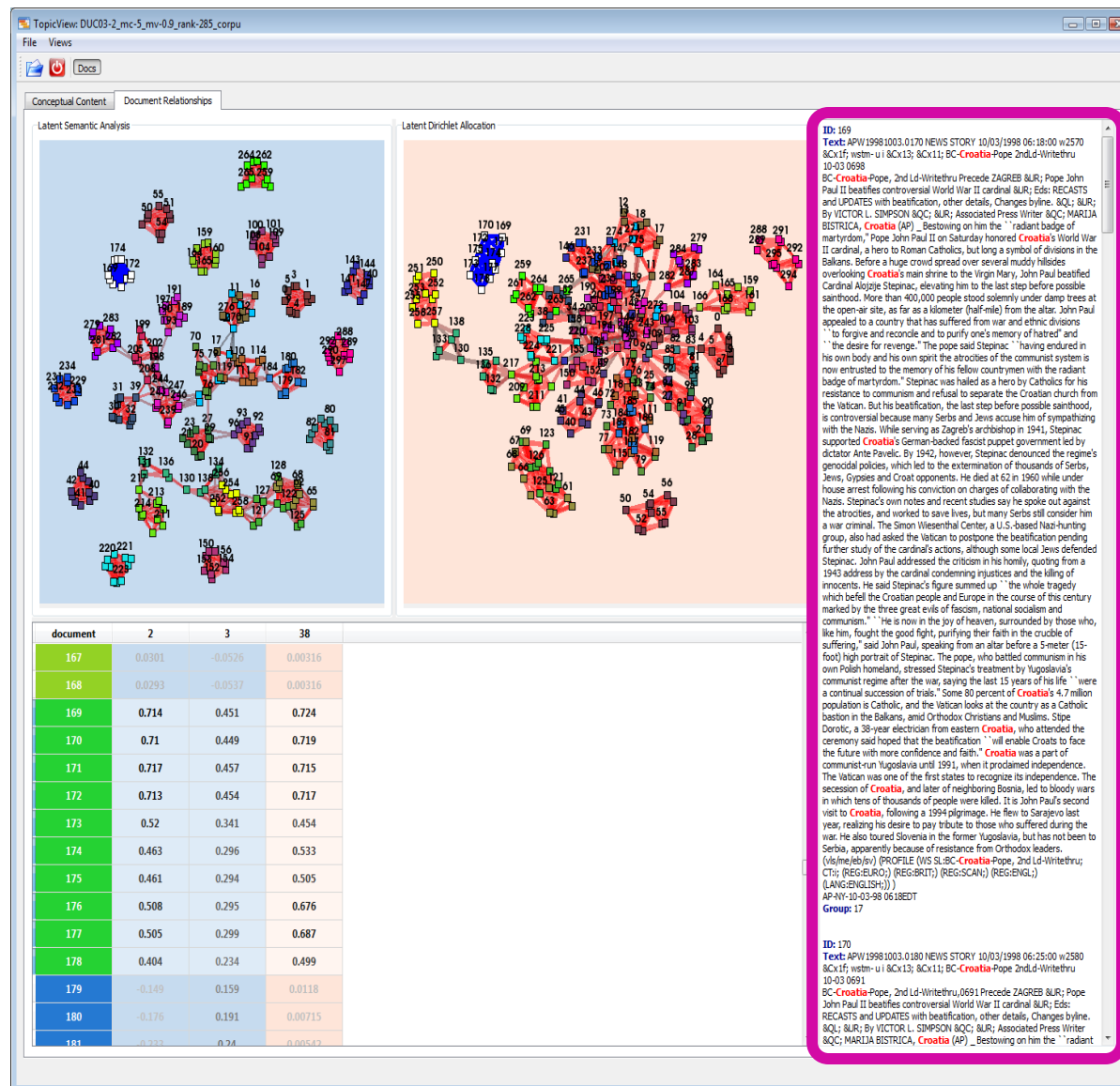
LSA
Document
Similarity
Graph



LDA
Document
Similarity
Graph

Document Topic Table



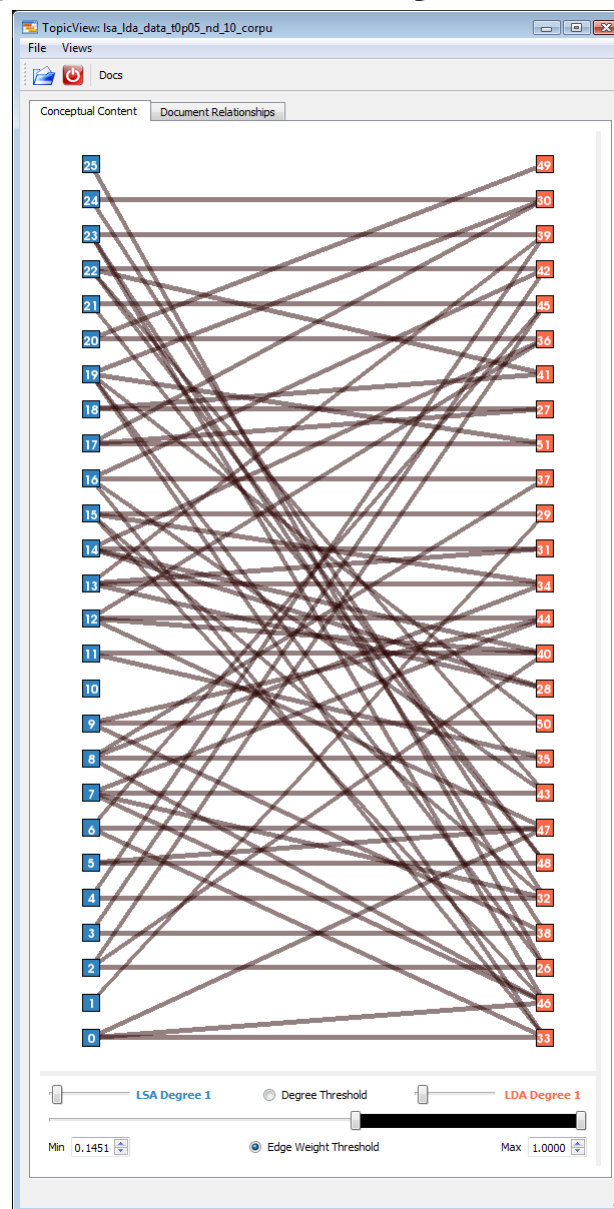
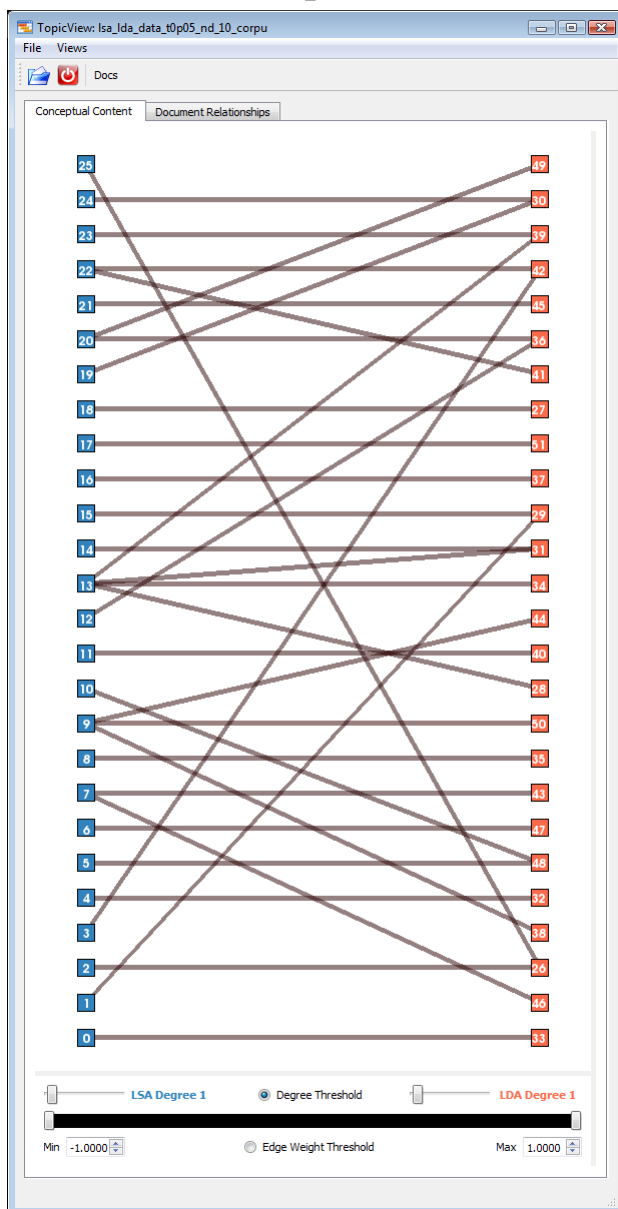


Alphabet Data Case Study

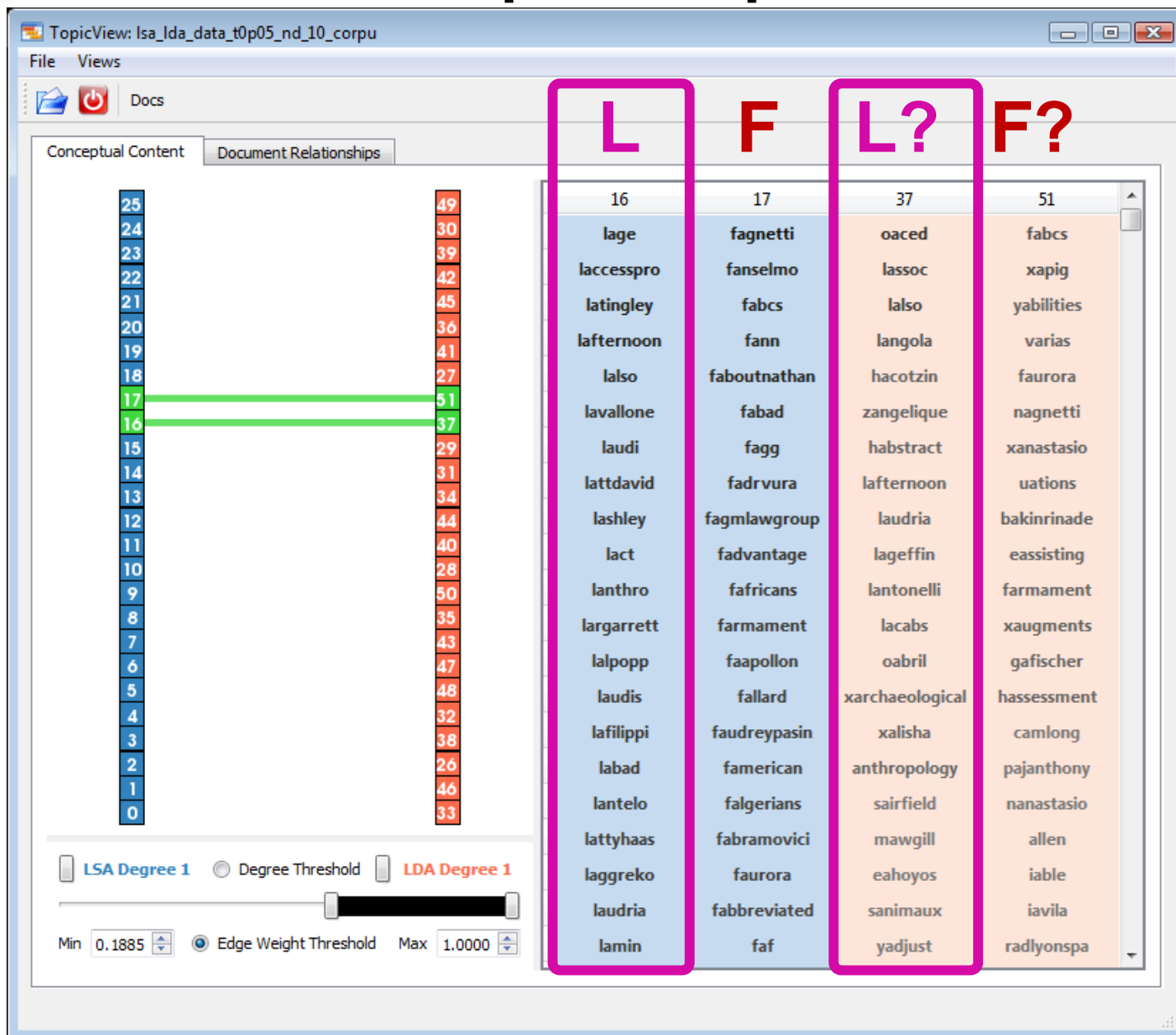
Synthetic Data for verification

- 26 clusters (one per letter), 10 documents each
- Each document contains only words starting with a single letter
 - absorbent autonomic appeals anthology aristocrats ...
 - bacquire bairbags baiming babomination battorney bafter ...
 - cadvisory cassumption cappears camount canthropology
 - ...
- Each algorithm given concept/topic count of 26

Alphabet Topic Similarity



Term/Topic Comparison



Document-Topic Weights

TopicView: Isa_lda_data_t0p05_nd_10_corpu

File Views

Conceptual Content Document Relationships

L **F** **L?** **F?**

document	16	17	37	51
107	-7.59e-10	1.66e-09	0.0387	0.0779
108	2.22e-10	-4.87e-10	0.0649	0.0289
109	-7.53e-11	1.66e-10	0.0681	0.0714
110	0.32	4.07e-10	0.0583	0.0616
111	0.393	-1.84e-09	0.124	0.042
112	0.363	-2.11e-09	0.0714	0.0387
113	0.353	3.7e-09	0.0812	0.0289
114	0.324	2.11e-09	0.0649	0.042
115	0.413	1.52e-10	0.0289	0.0224
116	0.366	9.47e-10	0.156	0.0322
117	0.392	-5.06e-10	0.0126	0.0289
118	0.349	-1.48e-09	0.0649	0.0191
119	0.4	-7.99e-10	0.0452	0.0158
120	-2.82e-10	6.23e-10	0.0224	0.0354
121	1.26e-09	-2.76e-09	0.0551	0.0126
122	-8.28e-10	1.81e-09	0.0452	0.0224
123	-1.22e-09	2.67e-09	0.042	0.0387

TopicView: Isa_lda_data_t0p05_nd_10_corpu

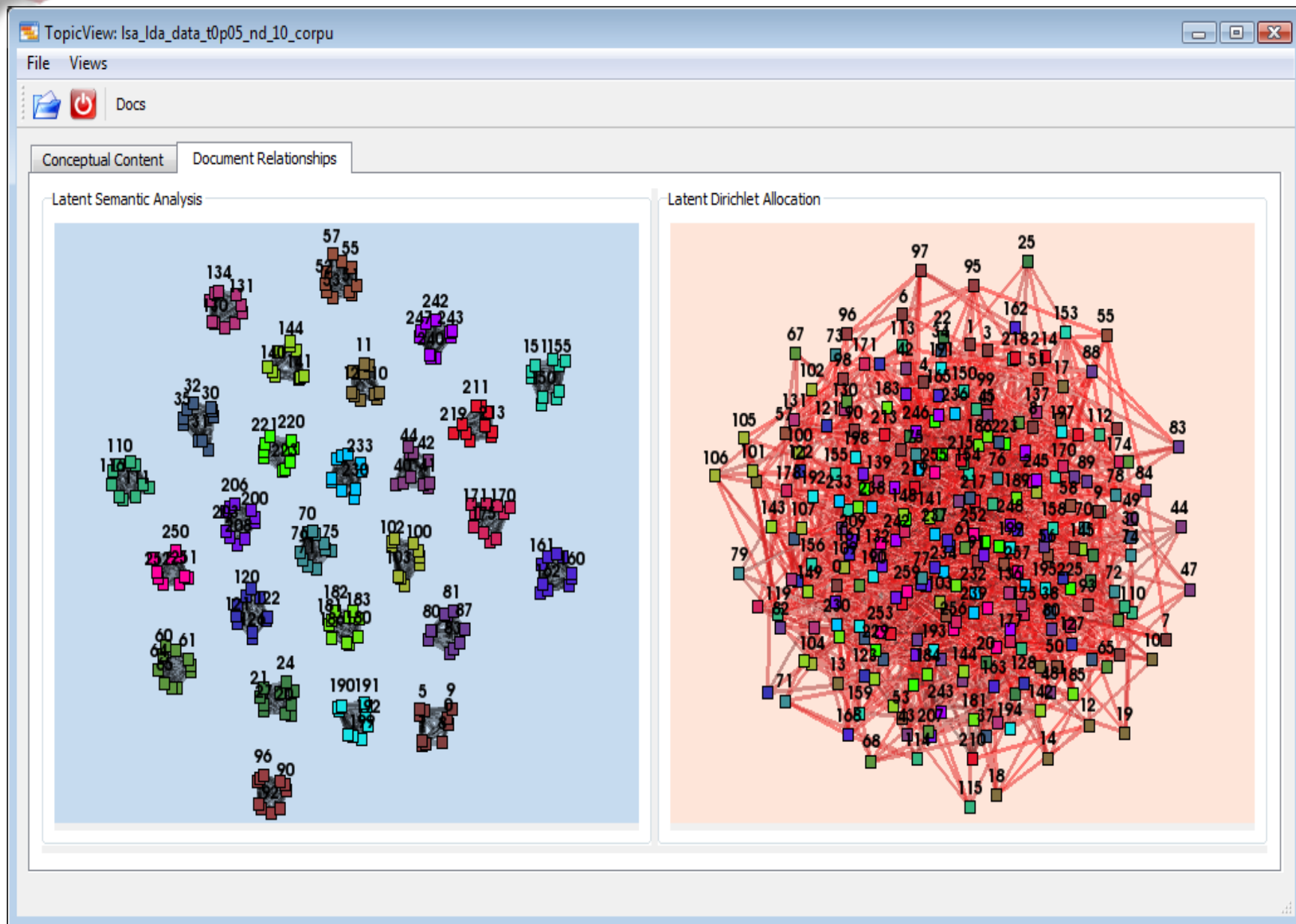
File Views

Conceptual Content Document Relationships

L **F** **L?** **F?**

document	16	17	37	51
47	-1.65e-09	3.62e-09	0.0256	0.0943
48	2.74e-09	-6.01e-09	0.0387	0.0191
49	-4.54e-10	9.89e-10	0.0681	0.0158
50	-1.89e-09	0.449	0.0256	0.124
51	3.84e-09	0.312	0.0747	0.0354
52	9.75e-11	0.367	0.0256	0.0583
53	-1.37e-09	0.38	0.0681	0.0779
54	7.54e-10	0.297	0.0225	0.0653
55	2.72e-09	0.319	0.042	0.0387
56	-1.21e-09	0.497	0.0354	0.0975
57	-7.96e-10	0.355	0.0256	0.0354
58	2.49e-10	0.341	0.0224	0.0747
59	-4.94e-10	0.316	0.0289	0.0256
60	1.08e-09	-2.36e-09	0.0616	0.0485
61	-1.45e-10	3.21e-10	0.0452	0.0583
62	-5.06e-10	1.11e-09	0.0158	0.0158
63	-4.22e-10	9.26e-10	0.0354	0.0224

Clustering Evaluation



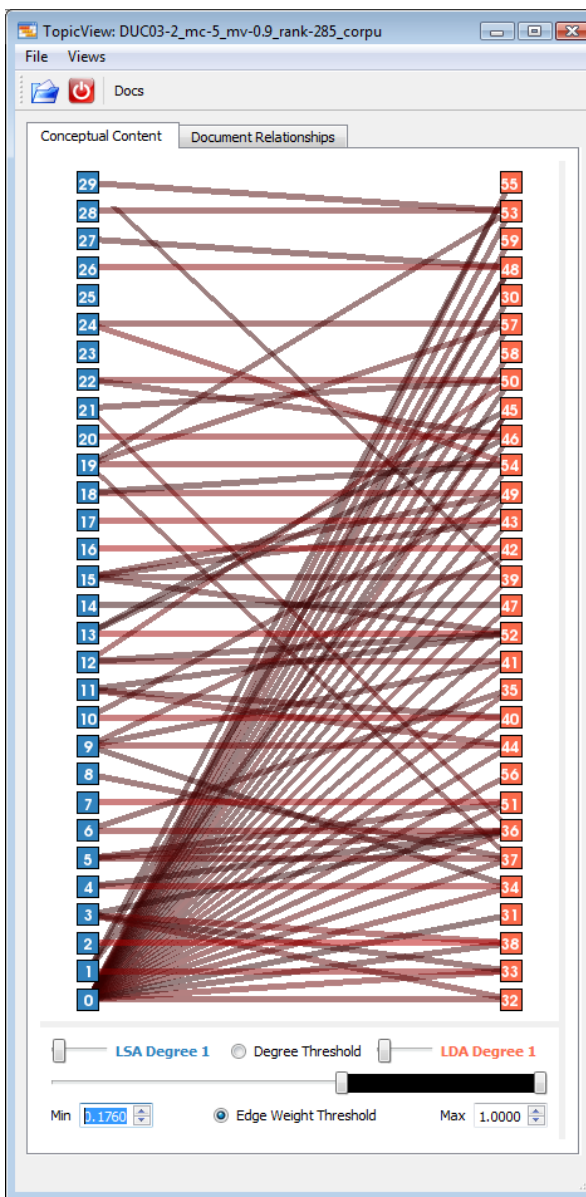
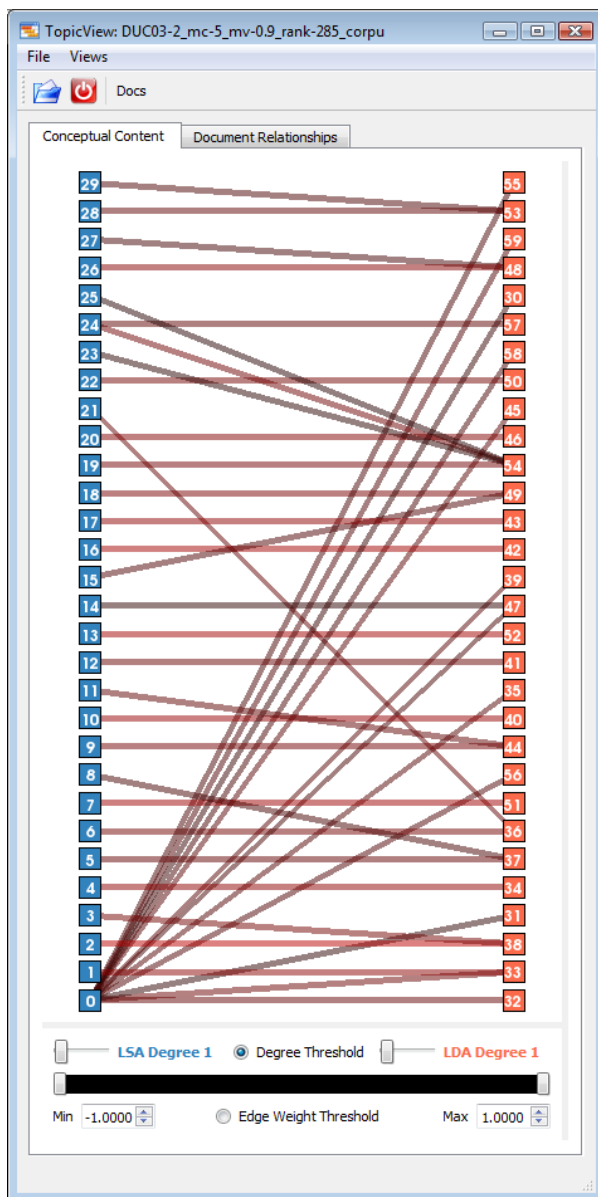


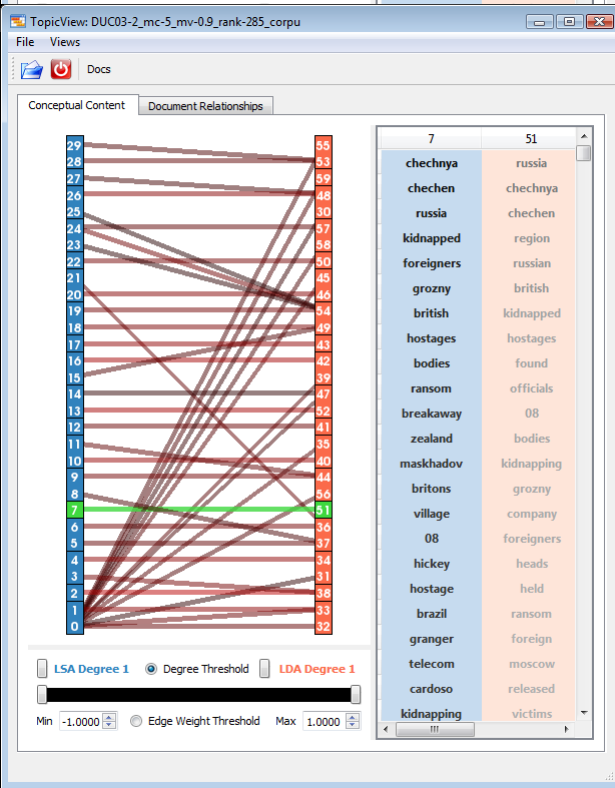
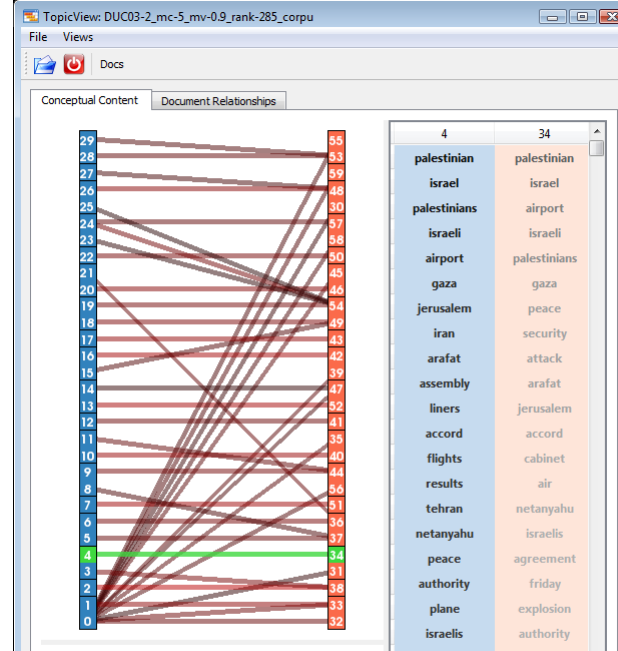
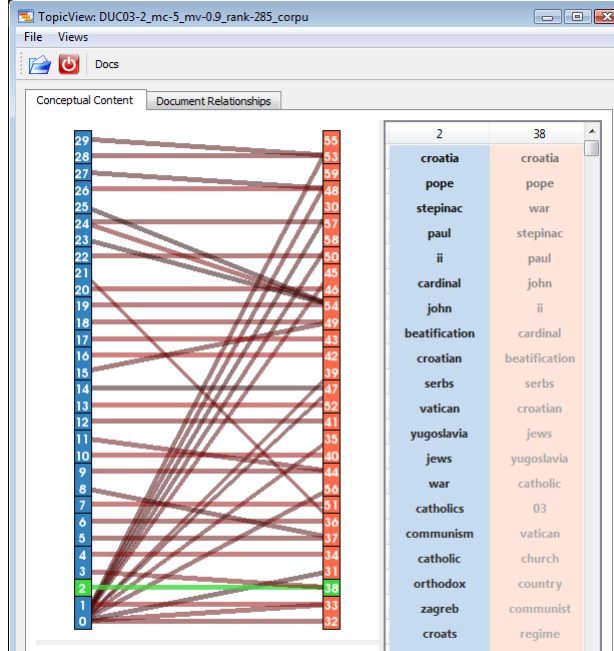
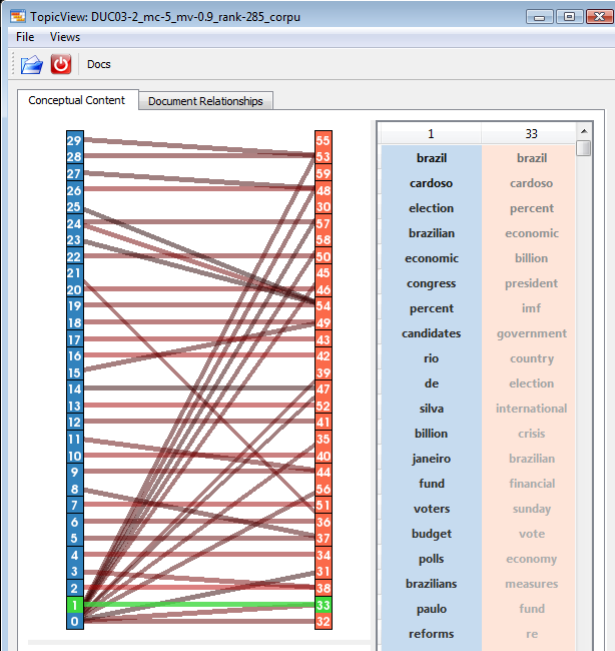
DUC Data Case Study

Document Understanding Conference (DUC) Data (real world)

- 30 clusters, ~10 documents each
- Human categorized around particular topic/event
- Associated Press articles
- New York Times articles
- Each algorithm given concept/topic count of 30

DUC Topic Similarity





LSA Combines Topics

TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

File Views

Conceptual Content Document Relationships

9	11	44
pinochet	pinochet	pinochet
spanish	fire	spanish
chile	spanish	chile
rights	chile	arrest
human	sweden	britain
palestinian	timor	british
chilean	britain	human
commission	dance	law
palestinians	chilean	london
britain	indonesia	rights

Pinochet
Arrest

Pinochet Arrest
Dance Hall Fire
Timor Unrest

Pinochet
Arrest

TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

File Views

Doc 121 connects Chile, Spanish, Fire

Conceptual Content Document Relationships

package has been estimated to total \$30 billion. Brazil will not collapse, he said, adding that Argentina and **Chile** were among the keenest to see the IMF reach agreement with their giant neighbor. Less upbeat about Russia, Camdessus said the Fund was still working toward a financing package, but that the government still lacked a program. This, he said, would take time, and appealed to governments to provide humanitarian aid in the interim. The IMF has put together bailout packages totaling over \$100 billion during what Camdessus called "the most uncertain year ever." He said the crisis has underlined the need for prevention, and the transformation of the Fund's primary role away from that of "the firefighter who arrives when the fire is already burning." Greater private sector involvement in the prevention and resolution of future economic crisis would help. Camdessus said "Now, the

TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

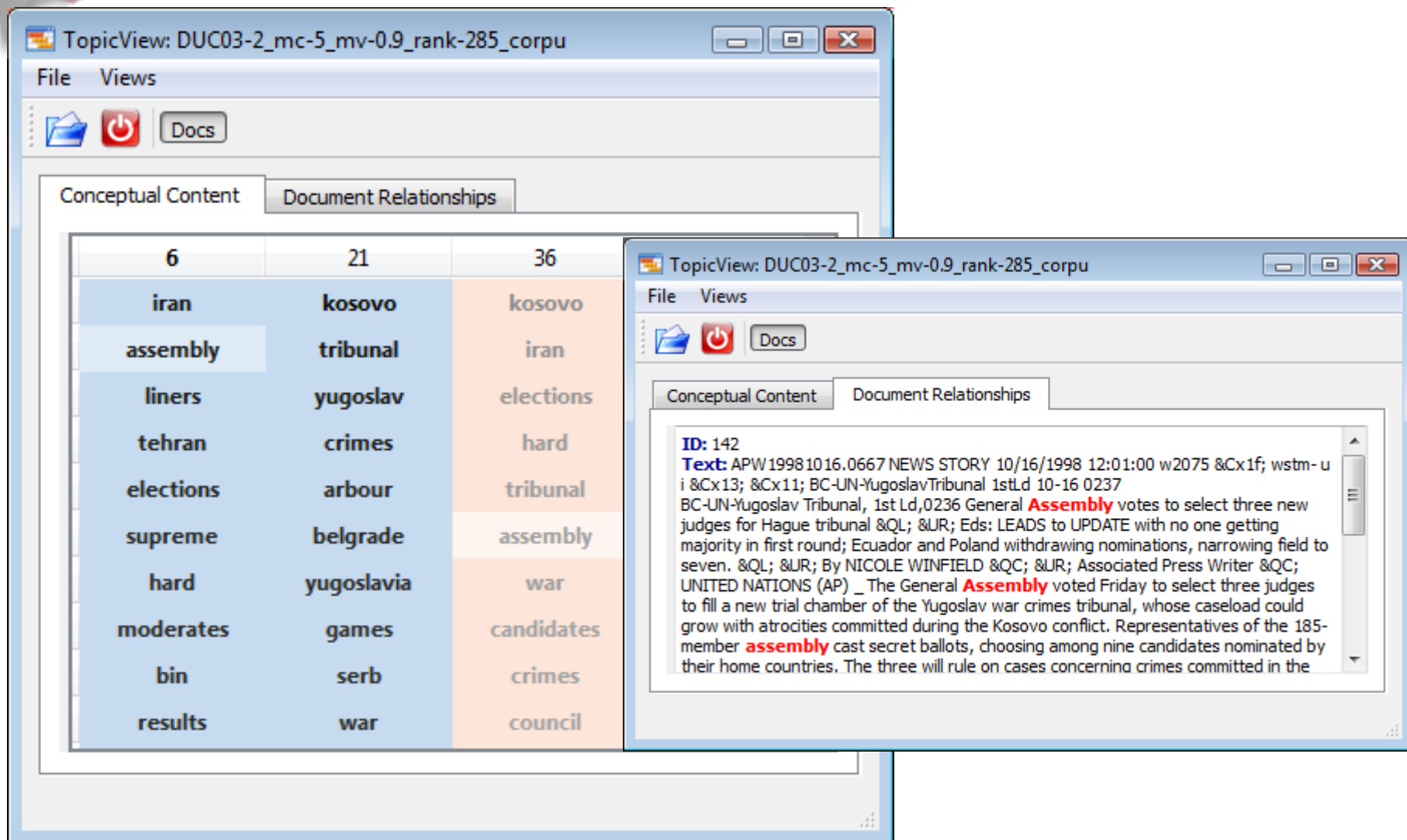
File Views

Doc 87 connects Pinochet & Timor

Conceptual Content Document Relationships

Indonesia annexed East Timor, a former Portuguese colony, after invading during a 1975 civil war that broke out when Portugal colonizers left. Indonesian troops have been accused of widespread abuses in East Timor, 1,200 miles (1,900 kilometers) east of Jakarta. Correia da Silva of Portugal's rightist Popular Party likens the bid to extradite Suharto with attempts in Britain to extradite former Chilean dictator Gen. Augusto **Pinochet** to Spain on charges that include genocide. **Pinochet** is in London awaiting a House of Lords panel decision on whether to extradite him. (PROFILE (WS SL:BC-Indonesia-Portugal; CT:i; (REG:EURO;) (REG:BRIT;) (REG:SCAN;) (REG:ENGL;) (LANG:ENGLISH;))) AP-NY-11-20-98 0533EST
Group: 8

LDA Combines Topics



TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

File Views

Conceptual Content Document Relationships

6	21	36
iran	kosovo	kosovo
assembly	tribunal	iran
liners	yugoslav	elections
tehran	crimes	hard
elections	arbour	tribunal
supreme	belgrade	assembly
hard	yugoslavia	war
moderates	games	candidates
bin	serb	crimes
results	war	council

TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

File Views

Conceptual Content Document Relationships

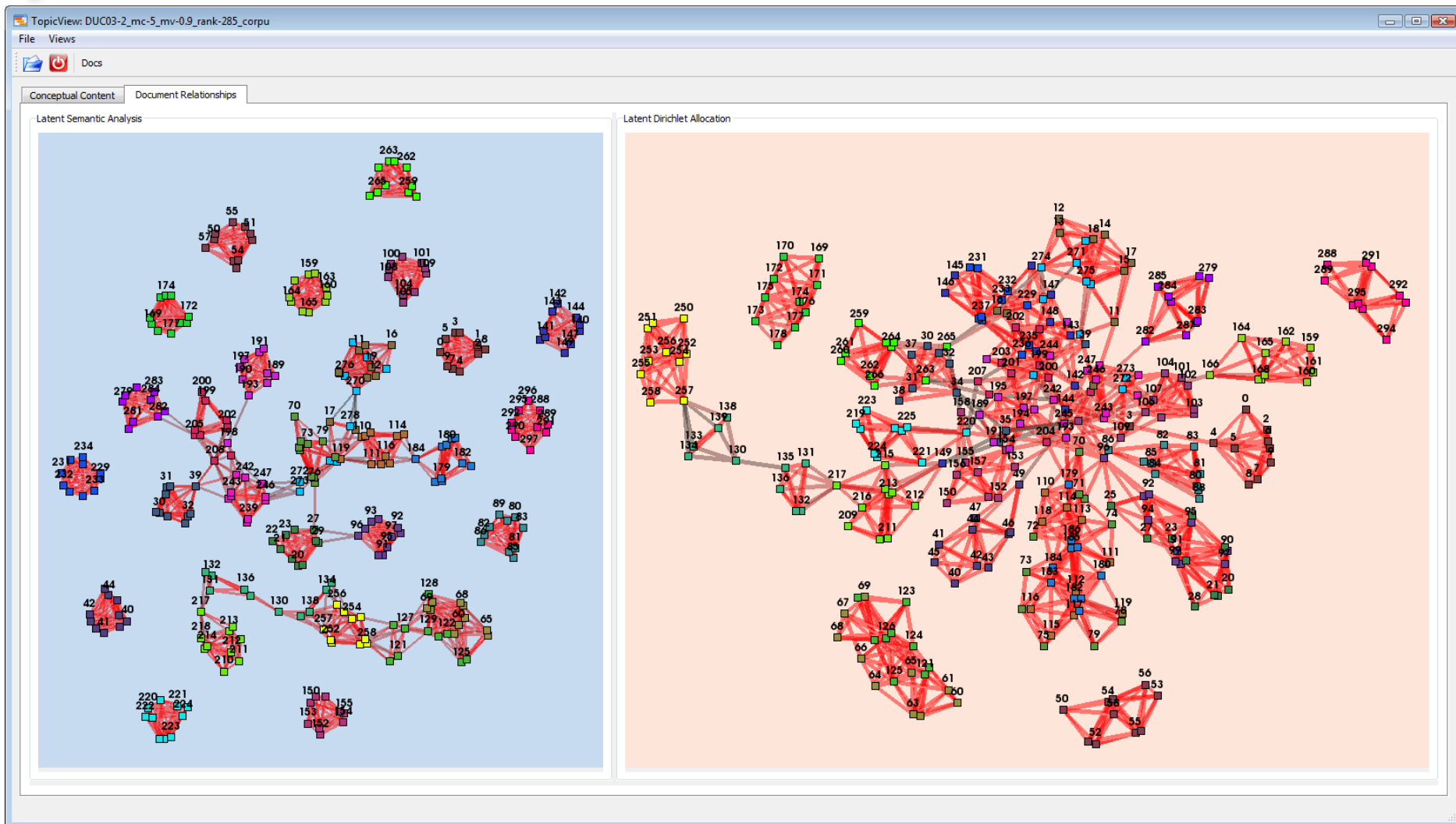
ID: 142
Text: APW19981016.0667 NEWS STORY 10/16/1998 12:01:00 w2075 &Cx1f; wstm- u i &Cx13; &Cx11; BC-UN-YugoslavTribunal 1stLd 10-16 0237
 BC-UN-Yugoslav Tribunal, 1st Ld,0236 General **Assembly** votes to select three new judges for Hague tribunal &QL; &UR; Eds: LEADS to UPDATE with no one getting majority in first round; Ecuador and Poland withdrawing nominations, narrowing field to seven. &QL; &UR; By NICOLE WINFIELD &QC; &UR; Associated Press Writer &QC; UNITED NATIONS (AP) _ The General **Assembly** voted Friday to select three judges to fill a new trial chamber of the Yugoslav war crimes tribunal, whose caseload could grow with atrocities committed during the Kosovo conflict. Representatives of the 185-member **assembly** cast secret ballots, choosing among nine candidates nominated by their home countries. The three will rule on cases concerning crimes committed in the

Iranian
Elections

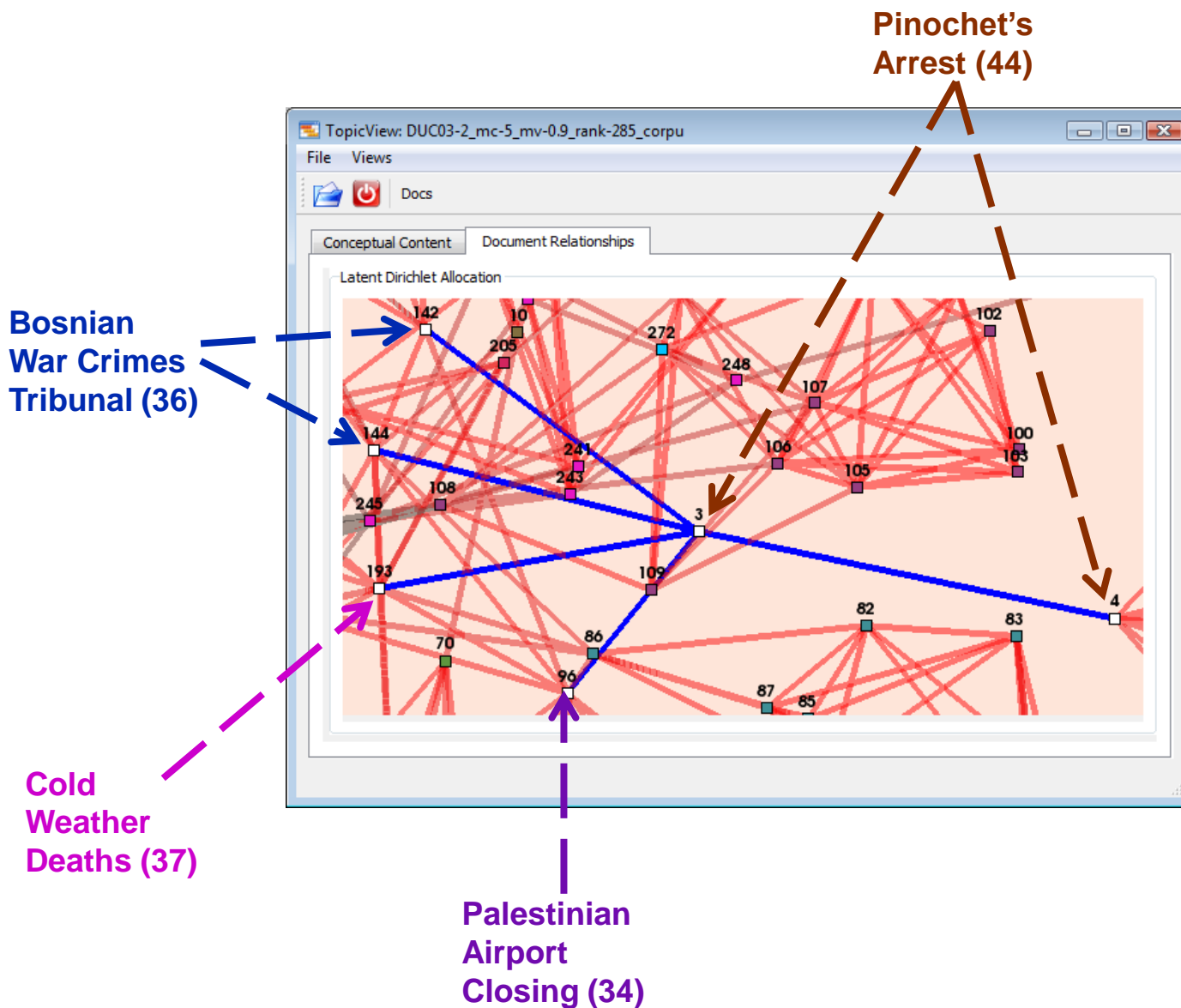
Bosnian
Tribunal

Iranian
Elections
Bosnian
Tribunal

DUC Document Relationships



LDA Unexpected Connections



Documents more strongly connected to Topic 30 than conceptual topics

TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

File Views

Conceptual Content Document Relationships

Latent Dirichlet Allocation

document	30	34	36	37	44
0	0.257	0.00569	0.0125	0.0091	0.501
1	0.241	0.00551	0.00424	0.00424	0.579
2	0.191	0.00524	0.00622	0.00917	0.6
3	0.387	0.00859	0.0112	0.00859	0.261
4	0.21	0.00815	0.0112	0.0143	0.464
5	0.0092	0.0023	0.00782	0.00506	0.622
6	0.01	0.00787	0.0122	0.00787	0.579
7	0.00857	0.00359	0.016	0.00857	0.539
8	0.00313	0.00407	0.00782	0.00407	0.454
9	0.00587	0.0103	0.0103	0.00808	0.525

TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

File Views

Conceptual Content Document Relationships

Latent Dirichlet Allocation

document	30	34	36	37	44
140	0.198	0.00412	0.12	0.00412	0.014
141	0.219	0.01	0.393	0.00729	0.0606
142	0.338	0.0121	0.193	0.0121	0.0464
143	0.272	0.0102	0.291	0.0102	0.0126
144	0.384	0.00998	0.0938	0.00998	0.0489
145	0.176	0.00687	0.601	0.00406	0.0078
146	0.182	0.0231	0.616	0.00791	0.00316
147	0.213	0.0162	0.348	0.00905	0.0705
148	0.239	0.00577	0.366	0.0161	0.0386
149	0.00753	0.0132	0.178	0.0156	0.0237

TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

File Views

Conceptual Content Document Relationships

Latent Dirichlet Allocation

document	30	34	36	37	44
91	0.143	0.71	0.00698	0.00363	0.0053
92	0.282	0.36	0.0202	0.00938	0.00938
93	0.321	0.404	0.0122	0.00644	0.0122
94	0.306	0.415	0.0143	0.0143	0.00843
95	0.242	0.484	0.0108	0.00907	0.00737
96	0.458	0.169	0.0166	0.0135	0.0135
97	0.00656	0.48	0.00529	0.0154	0.00719
98	0.00955	0.692	0.00645	0.0049	0.00335
99	0.00711	0.696	0.00788	0.00483	0.0033
100	0.171	0.0372	0.0109	0.00437	0.0241

TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

File Views

Conceptual Content Document Relationships

Latent Dirichlet Allocation

document	30	34	36	37	44
190	0.247	0.00617	0.00902	0.496	0.00902
191	0.375	0.00765	0.00994	0.313	0.00765
192	0.424	0.01	0.01	0.269	0.01
193	0.384	0.0142	0.0115	0.115	0.0115
194	0.342	0.00896	0.00896	0.345	0.0116
195	0.278	0.0147	0.0147	0.35	0.00916
196	0.363	0.00936	0.00936	0.276	0.015
197	0.339	0.00782	0.0172	0.318	0.00782
198	0.0108	0.00789	0.00641	0.533	0.00493
199	0.189	0.0093	0.00372	0.568	0.0115

Topic 30 - AP wire source

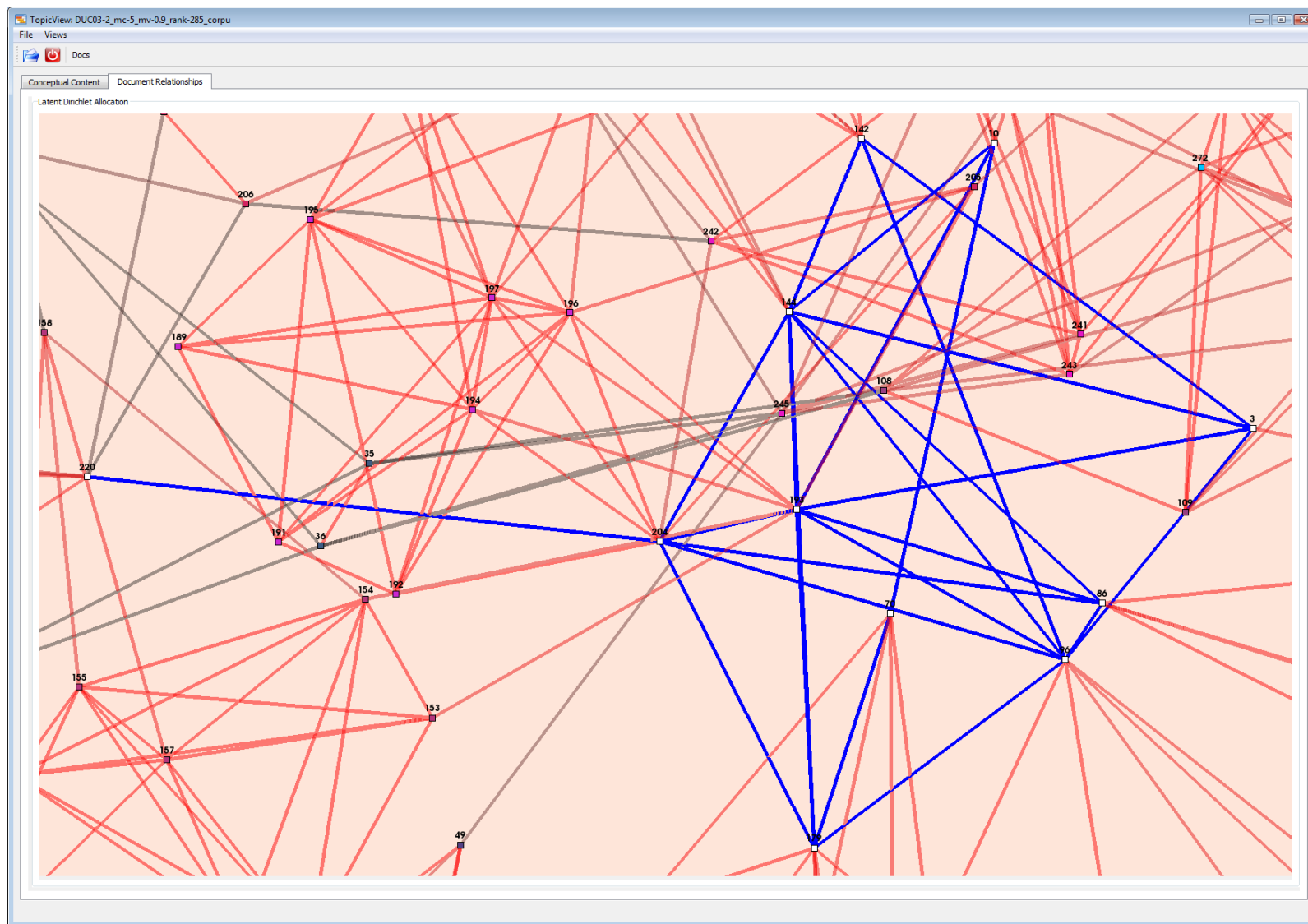
TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu

File Views

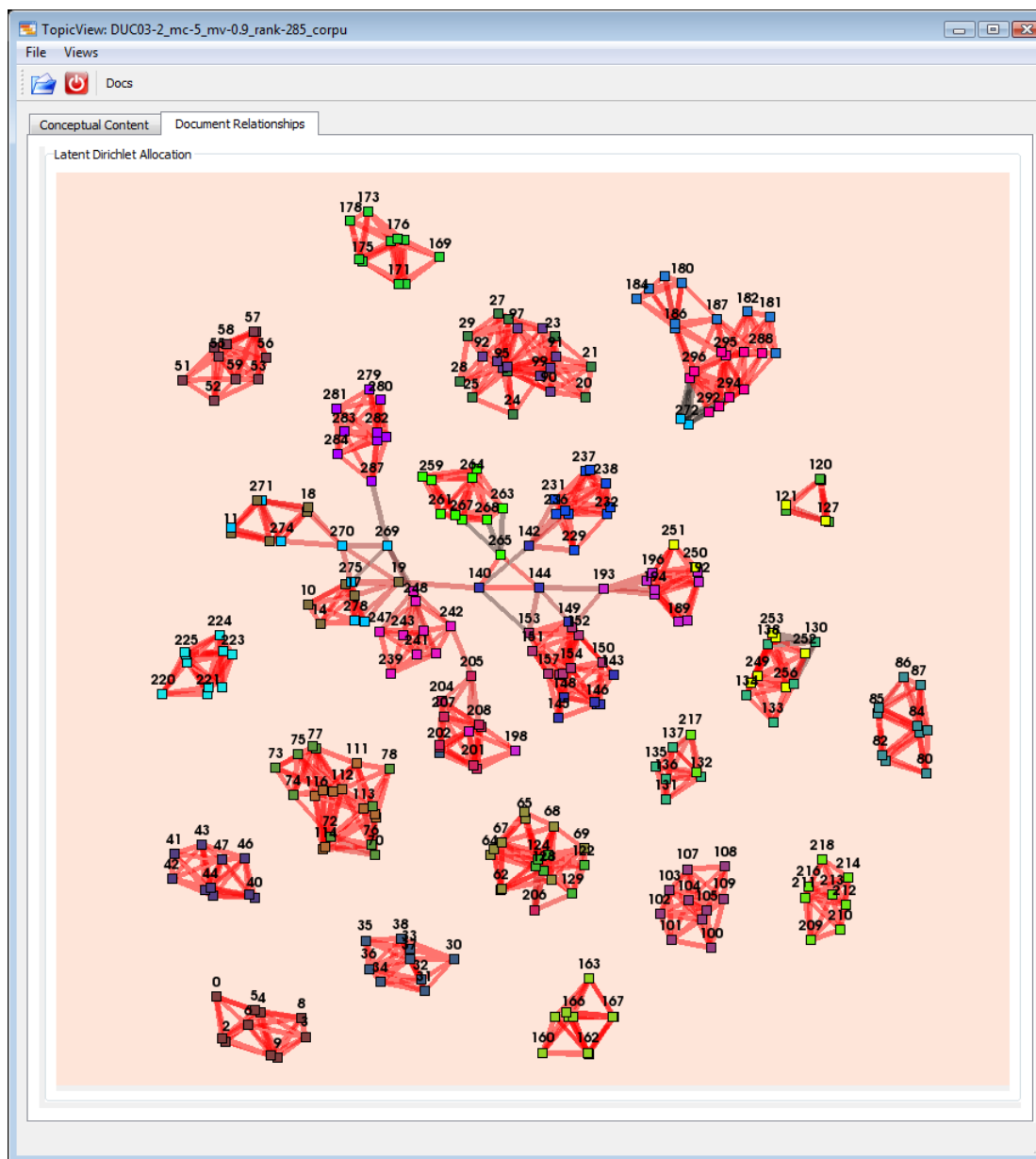
Conceptual Content Document Relationships

30	34	36	37	44
reg	palestinian	kosovo	cold	pinochet
bc	israel	iran	poland	spanish
ap	airport	elections	typhoon	chile
ur	israeli	hard	philippines	arrest
date	palestinians	tribunal	weather	britain
body	gaza	assembly	death	british
trailer	peace	war	hit	human
header	security	candidates	air	law
slug	attack	crimes	northern	london
doc	arafat	council	babs	rights
headline	jerusalem	leader	toll	international
docno	accord	yugoslavia	died	chilean
text	cabinet	tehran	minus	argentina

Bridging documents: conceptual content outweighed by source content



LDA rerun without header tags



Conclusions

- LSA concepts provide good summarizations over broad document groups
- LDA topics are focused on smaller groups
- LDA's limited groups and probabilistic mechanism provides better labeling
- LSA's document relationships do not include extraneous connections between disparate topics
- Better graphs
- Better labels